CG020 Genomika

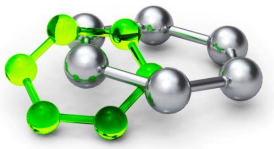
Přednáška 1

Úvod do bioinformatiky

Jan Hejátko

Funkční genomika a proteomika rostlin,
Středoevropský technologický institut (CEITEC)
a
Národní centrum pro výzkum biomolekul,
Přírodovědecká fakulta,

Masarykova univerzita, Brno
hejatko@sci.muni.cz, www.ceitec.eu



Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

Schéma předmětu

- **Kapitola 01**
  - **Úvod do bioinformatiky**
- **Kapitola 02**
  - **Identifikace genů**
- **Kapitola 03**
  - **Přístupy reverzní genetiky**
- **Kapitola 04**
  - **Přístupy genetiky přímé**

Schéma předmětu

- **Kapitola 05**
  - **RNA interference a genomové editování**
- **Kapitola 06**
  - **Genová exprese a chemická genetika**
- **Kapitola 07**
  - **Protein-proteinové interakce a jejich analýza**
- **Kapitola 08**
  - **Současné metody sekvenování DNA**

Schéma předmětu

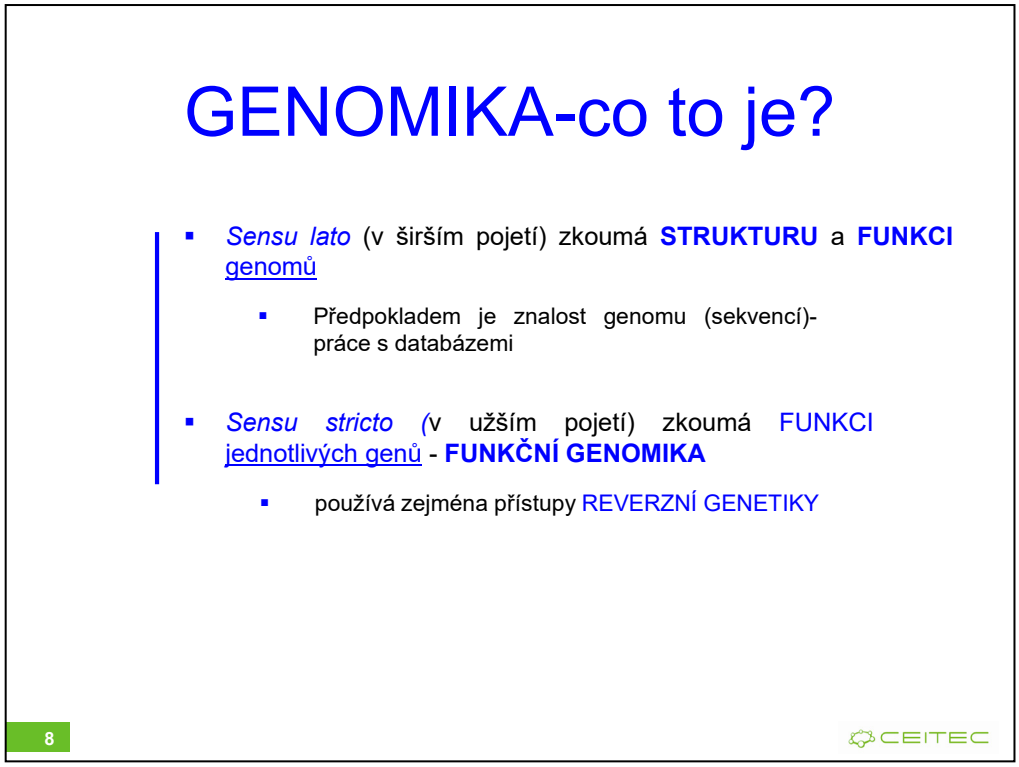
- **Kapitola 09**
  - **Struktura genomů**
- **Kapitola 10**
  - **Evoluce genomů**
- **Kapitola 11**
  - **Genomika a systémová biologie**
- **Kapitola 12**
  - **Praktické aspekty funkční genomiky**
  - **Modelové organismy**
  - **PCR**

Literatura

- Literární zdroje pro kapitolu 01:
  - **Bioinformatics and Functional Genomics**, 3rd Edition, Jonathan Pevsner, Wiley-Blackwell, 2015 http://www.bioinfbook.org/php/?q=book3
  - **Úvod do praktické bioinformatiky**, Fatima Cvrčková, 2006, Academia, Praha
  - **Plant Functional Genomics**, ed. Erich Grotewold, 2003, Humana Press, Totowa, New Jersey

Osnova

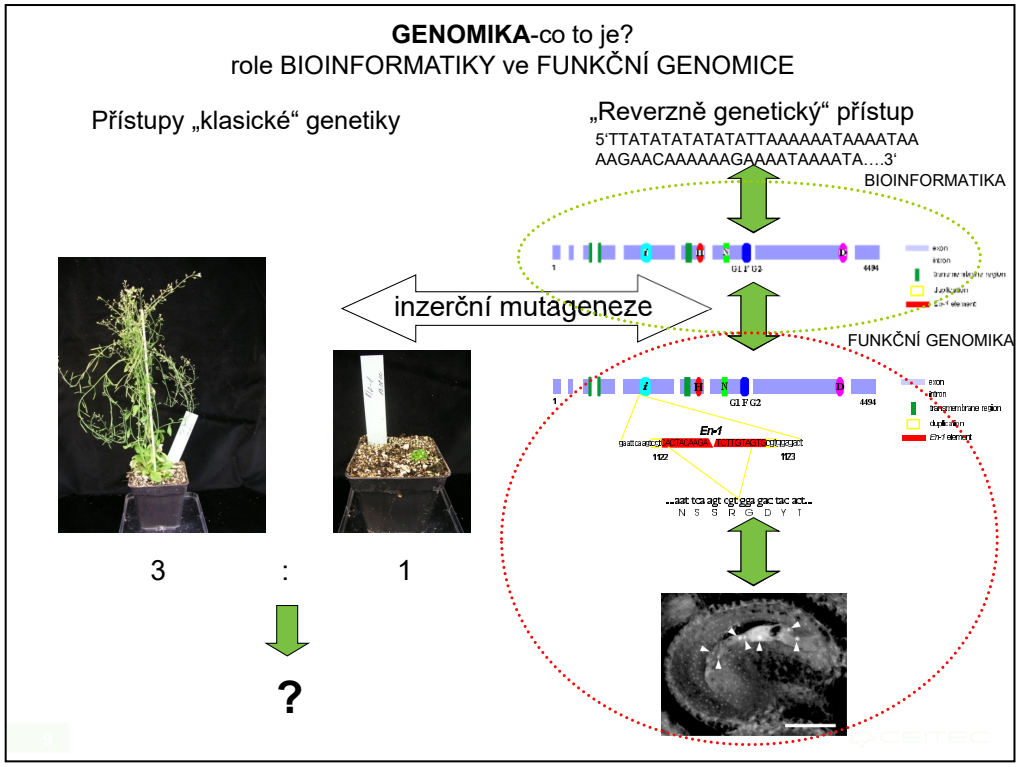
- Schéma předmětu
- Definice



Genomics is a science discipline that is interested in the analysis of genomes. Genome of each organism is a complex of all genes of the respective organism. The genes could be located in cytoplasm (prokaryots) nucleus (in most euckaryotic organisms), mitochondria or chloroplasts (in plants).

The critical prerequisite of genomics is the knowledge of gene sequences.

Functional genomics is interested in function of individual genes.



With the knowledge of gene sequences (or the knowledge of the gene files in the individual organisms, i.e. the knowledge of genomes), **Reverse Genetics** appears that allows study their function.

In comparison to ”classical” or **Forward Genetics**, starting with the phenotype, the reverse genetics starts with the sequence identified as a gene in the sequenced genome. The gene identification using approaches of **Bioinformatics** will be described later (see Lesson 02).

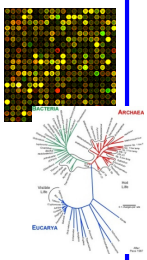
Reverse genetics uses a spectrum of approaches that will be described in the Lesson 03 that allow isolation of sequence-specific mutants and thus their phenotype analysis.

The necessity of having phenotype alterations in the forward genomics approach introduces important difference between those two approaches. Thus, the gene is no longer understood as a factor (*trait*) determining *phenotype,* but rather as a piece of DNA characterized by the unique *string of nucleotides*. i.e. **physical DNA molecule**.

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY

Bioinformatika



- **Definice bioinformatiky** (podle NIH vědeckého a technologického konsorcia pro biomedicínské informace)

Výzkum, vývoj nebo aplikace výpočetních nástrojů a přístupů za účelem zvyšování rozvoje využití biologických, lékařských, dat o chování nebo zdraví, včetně těch, které umožňují taková data získávat, ukládat, organizovat, archivovat, analyzovat nebo vizualizovat.

CEITEC

NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY
July 17, 2000

The following working definition of bioinformatics and computational biology were developed by the BISTIC Definition Committee and released on July 17, 2000. The committee was chaired by Dr. Michael Huerta of the National Institute of Mental Health and consisted of the following members:

Bioinformatics Definition Committee BISTIC Members Expert Members
Michael Huerta (Chair) Gregory Downing
Florence Haseltine Belinda Seto
Yuan Liu

Preamble
Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Definition
The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

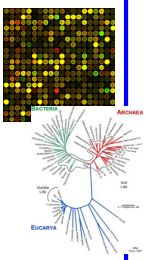
What is Bioinformatics?

- Interface of **biology** and **computers**
- Analysis of **proteins, genes** and **genomes** using **computer algorithms** and **computer databases**
- **Genomics** is the **analysis of genomes**. The **tools of bioinformatics** are used **to make sense** of the **billions** of **base pairs of DNA** that are sequenced by genomics projects.

J. Pevsner,
http://www.bioinfbook.org/index.php

CEITEC

Bioinformatika



- **Bioinformatika** ve **funkční genomice**
  - **Zpracování a analýza sekvenačních dat**
    - Identifikace referenčních sekvencí
    - Identifikace genů
    - Identifikace homologů, ortologů a paralogů
    - Korelační analýzy mezi genomy a fenotypy (včetně člověka)
  - **Zpracování a analýza transkripčních dat**
    - Transkripční profilování pomocí DNA čipů nebo next-gen sekvenování
  - **Vyhodnocování experimentálních dat** a **predikce nových regulací** v přístupech **systémové biologie**
    - Matematické modelování genových regulačních sítí

CEITEC

Množství informace v DNA

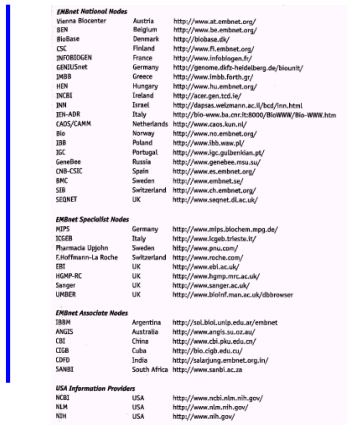
- Every **bp**= **4 bits**
- Human genome = ~**3 billion bp**
  - = 4 × 3 × 10^9
  - =1.2 × 10^{10} bits
  - =1.5 × 10^9 bytes (**1.5 GB**)
- This amount of information is contained in a cell nucleus with 10µm diameter

- There is ~**2m** of DNA in **every somatic human cell**
  - **Each human** in composed of about **10^{12} cells**
  - Thus every human contains 2 × 1 **10^{12}** of DNA
    - **=2 × 10^9km** of DNA
  - Distance from the sun to Uranus = 2.8 × 109km
  - Each single human contains enough DNA to stretch from the sun to Uranus

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů

Spektrum on-line zdrojů



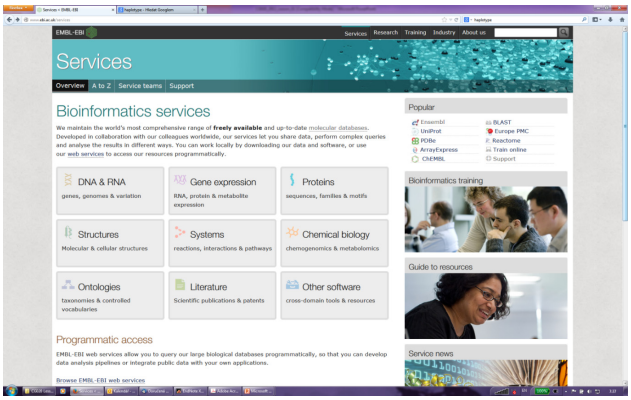
EMBnet National Nodes		
Vienna Biocenter	Austria	http://www.at.embnet.org/
BEN	Belgium	http://www.be.embnet.org/
BioBase	Denmark	http://biobase.dk/
CSC	Finland	http://www.fi.embnet.org/
INFOBIOGEN	France	http://www.infobiogen.fr/
GENIUSnet	Germany	http://genome.dkfz-heidelberg.de/biounit/
IMBB	Greece	http://www.imbb.forth.gr/
HEN	Hungary	http://www.hu.embnet.org/
INCBI	Ireland	http://acer.gen.tcd.ie/
INH	Israel	http://depsas.weizmann.ac.il/bcd/inn.html
IEN-ADR	Italy	http://bio-www.ba.cnr.it:8000/BioWWW/Bio-WWW.htm
CAOS/CAMM	Netherlands	http://www.caos.kun.nl/
Bio	Norway	http://www.no.embnet.org/
IBB	Poland	http://www.ibb.waw.pl/
IGC	Portugal	http://www.igc.gulbenkian.pt/
GeneBee	Russia	http://www.genebee.msu.su/
CNB-CSIC	Spain	http://www.es.embnet.org/
BMC	Sweden	http://www.embnet.se/
SIB	Switzerland	http://www.ch.embnet.org/
SEQNET	UK	http://www.seqnet.dl.ac.uk/
EMBnet Specialist Nodes		
MIPS	Germany	http://www.mips.biochem.mpg.de/
ICGEB	Italy	http://www.icgeb.trieste.it/
Pharmacia Upjohn	Sweden	http://www.pnu.com/
F.Hoffmann-La Roche	Switzerland	http://www.roche.com/
EBI	UK	http://www.ebi.ac.uk/
HGMP-RC	UK	http://www.hgmp.mrc.ac.uk/
Sanger	UK	http://www.sanger.ac.uk/
UMBER	UK	http://www.biolnf.man.ac.uk/dbbrowser
EMBnet Associate Nodes		
IBBM	Argentina	http://sol.bioi.unlp.edu.ar/embnet
ANGIS	Australia	http://www.angis.su.oz.au/
CBI	China	http://www.cbi.pku.edu.cn/
CIGB	Cuba	http://bio.cigb.edu.cu/
CDFD	India	http://talarjung.embnet.org.in/
SANBI	South Africa	http://www.sanbi.ac.za
USA Information Providers		
NCBI	USA	http://www.ncbi.nlm.nih.gov/
NLM	USA	http://www.nlm.nih.gov/
NIH	USA	http://www.nih.gov/

CEITEC

There are many of on-line resources that could be used.

Spektrum on-line zdrojů

- EBI http://www.ebi.ac.uk/services



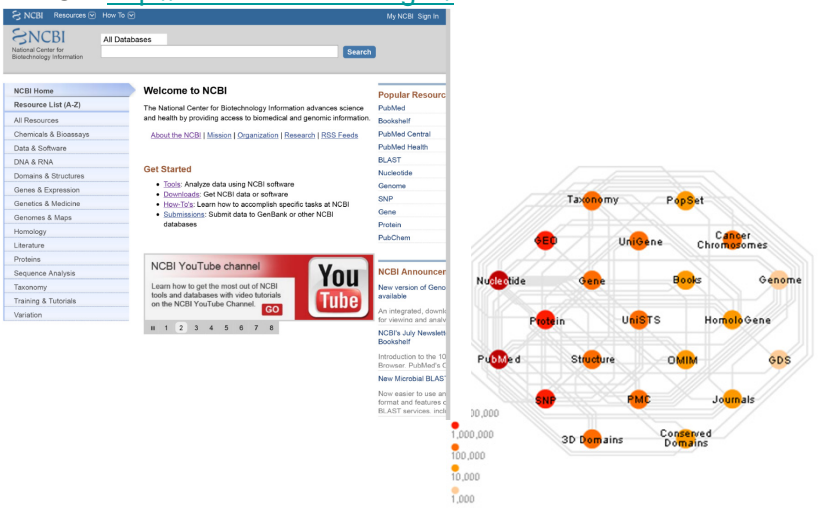
17

CEITEC

Nowadays, the resources are interconnected and could be accessed via dedicated web pages. Among the best and mostluy used www resources integrating plenty of database resources belong www portal of European Bioinformatics Institute (EBI) in Europe (Germany) and National Center of Biotechnology Information (NCBI) in the USA (

Spektrum on-line zdrojů

- NCBI http://www.ncbi.nlm.nih.gov/



18

CEITEC

Nowadays, the resources are interconnected and could be accessed via dedicated web pages.

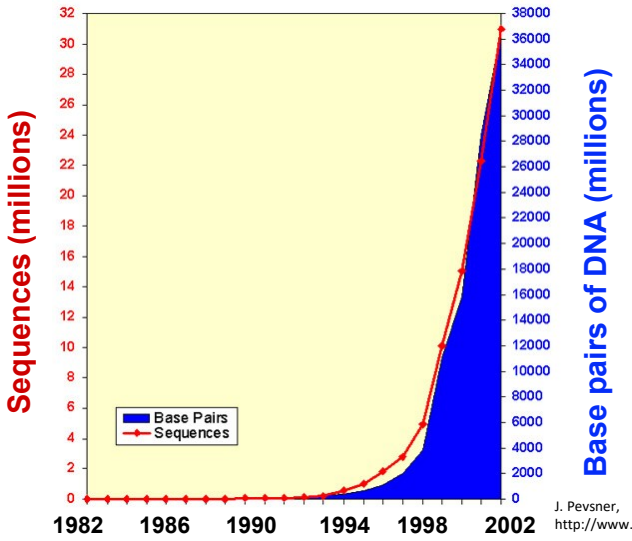
Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze

Primární databáze

- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - Sekvence v databázích tzv. „Velké trojky“:
    - **EMBL**
      - http://www.ebi.ac.uk/embl/
    - **GenBank,**
      - https://www.ncbi.nlm.nih.gov/
    - **DDBJ,**
      - http://www.ddbj.nig.ac.jp
  - denně vzájemná výměna a zálohování dat
  - velká datová náročnost (kapacita i software)

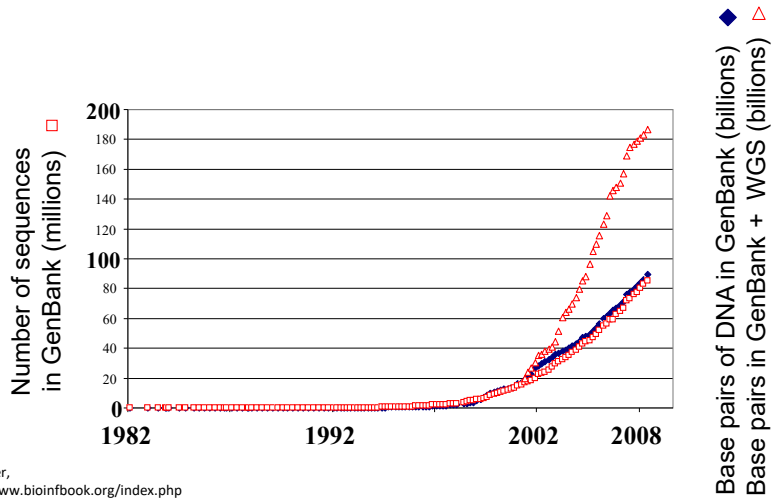
Growth of GenBank



J. Pevsner,
http://www.bioinfbook.org/index.php

CEITEC

Growth of GenBank + Whole Genome Shotgun (1982-November 2008): we reached 0.2 terabases

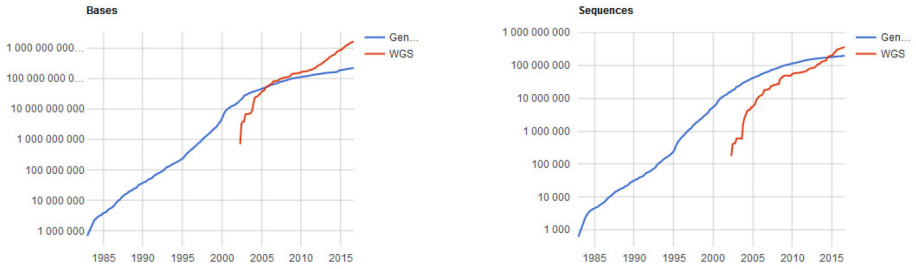


J. Pevsner,
http://www.bioinfbook.org/index.php

CEITEC

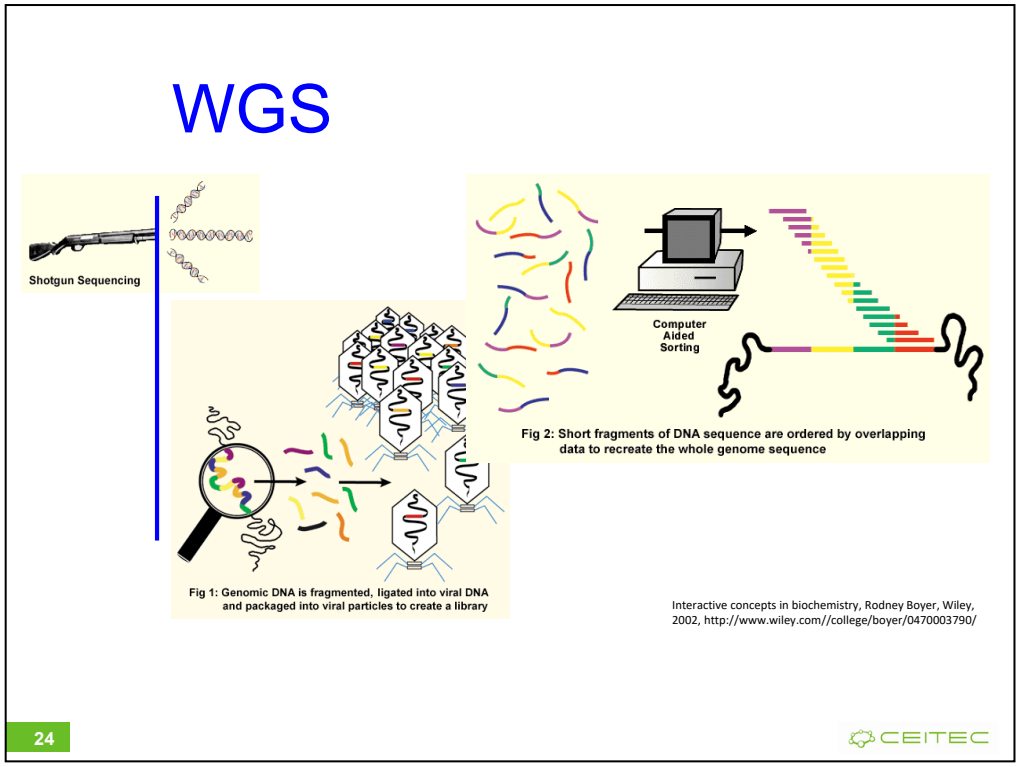
Growth of GenBank

Aug 2016



- Prosinec **1982** 680 338 bp, 606 sekvencí
- Duben **2002** 19 x 10^9 bp, 17 x 10^6 sekvencí + WGS 692 x 10^6 bp, 172 768 sekvencí
- Srpen **2016** 218 x 10^9 bp, 196 x 10^6 sekvencí + WGS 1,6 10^{12} bp, 360 x 10^6 sekvencí

CEITEC

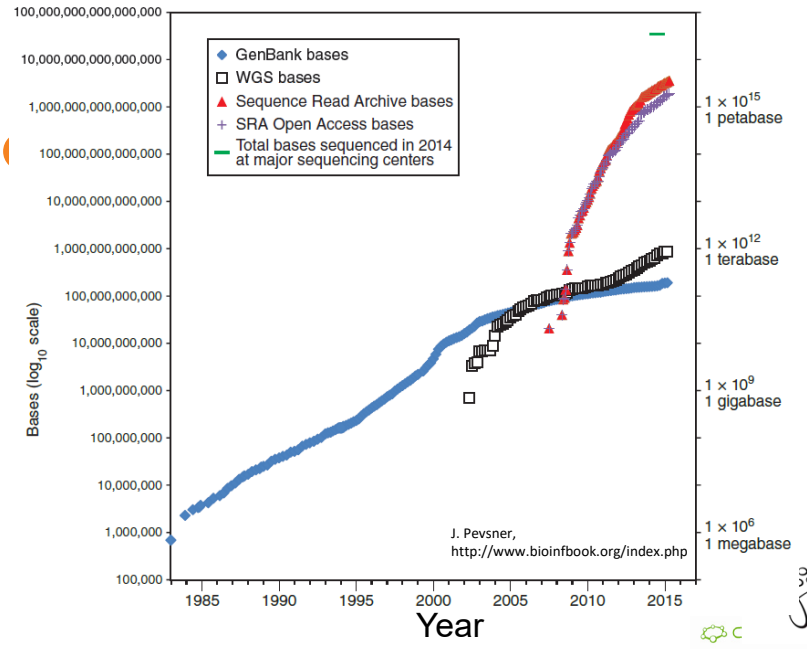


Shotgun sequencing allows a scientist to rapidly determine the sequence of very long stretches of DNA. The key to this process is fragmenting of the genome into smaller pieces that are then sequenced side by side, rather than trying to read the entire genome in order from beginning to end. The genomic DNA is usually first divided into its individual chromosomes. Each chromosome is then randomly broken into small strands of hundreds to several thousand base pairs, usually accomplished by mechanical shearing of the purified genetic material. Each of the short DNA pieces is then inserted into a DNA vector (a viral genome), resulting in a viral particle containing "cloned" genomic DNA (Fig. 1).

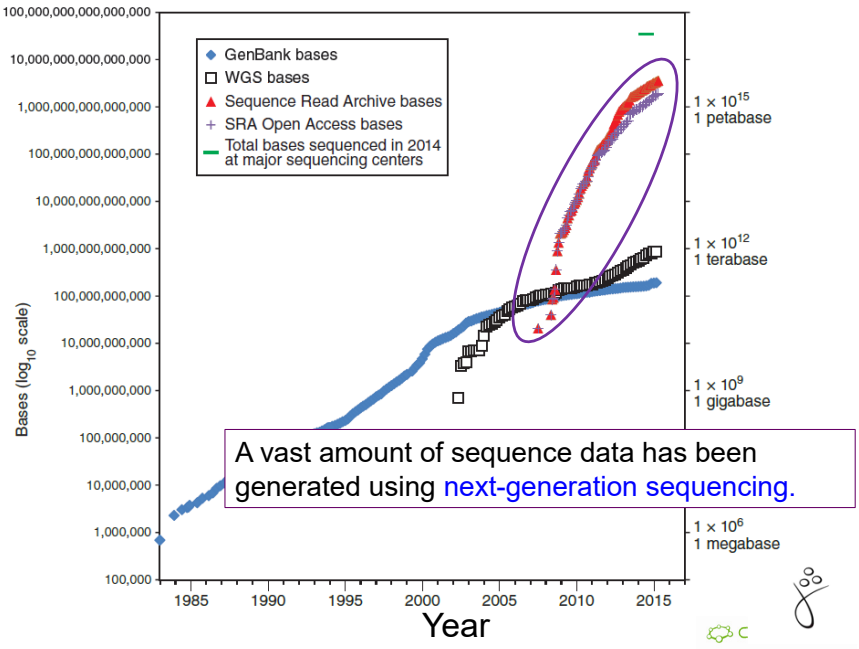
The collection of all the viral particles with all the different genomic DNA pieces is referred to as a library. Just as a library consists of a set of books that together make up all of human knowledge, a genomic library consists of a set of DNA pieces that together make up the entire genome sequence. Placing the genomic DNA within the viral genome allows bacteria infected with the virus to faithfully replicate the genomic DNA pieces. Additionally, since a little bit of known sequence is needed to start the sequencing reaction, the reaction can be primed off the known flanking viral DNA.

In order to read all the nucleotides of one organism, millions of individual clones are sequenced. The data is sorted by computer, which compares the sequences of all the small DNA pieces at once (in a "shotgun" approach) and places them in order by virtue of their overlapping sequences to generate the full-length sequence of the genome (Fig. 2). To statistically ensure that the whole genome sequence is acquired by this method, an amount of DNA equal to five to ten times the length of the genome must be sequenced. (Interactive concepts in biochemistry, Rodney Boyer, Wiley, 2002, http://www.wiley.com//college/boyer/0470003790/)

Growth of DNA Sequence in Repositories

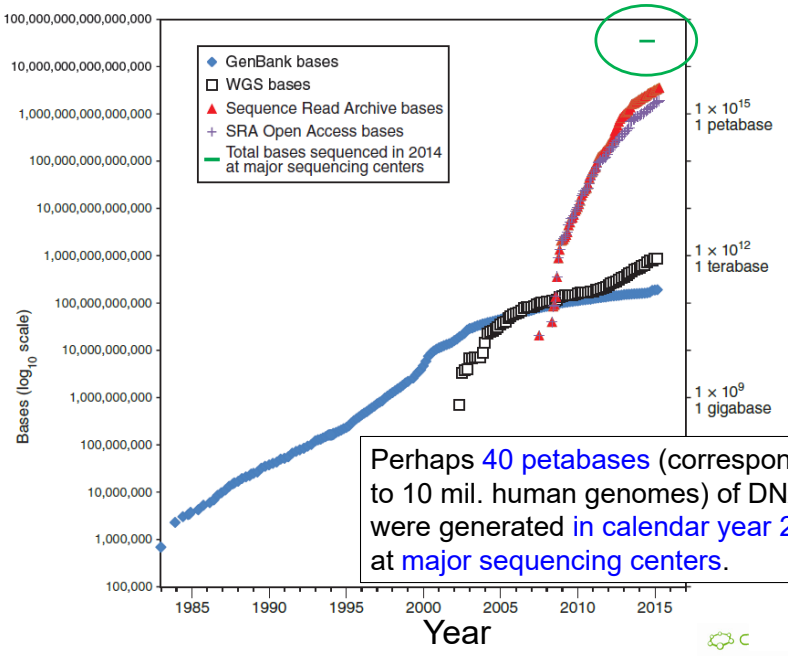


Growth of DNA Sequence in Repositories



A vast amount of sequence data has been generated using next-generation sequencing.

Growth of DNA Sequence in Repositories



Perhaps 40 petabases (corresponding to 10 mil. human genomes) of DNA were generated in calendar year 2014 at major sequencing centers.

B&FG 3e
Fig. 2-3
Page 22

Primární databáze

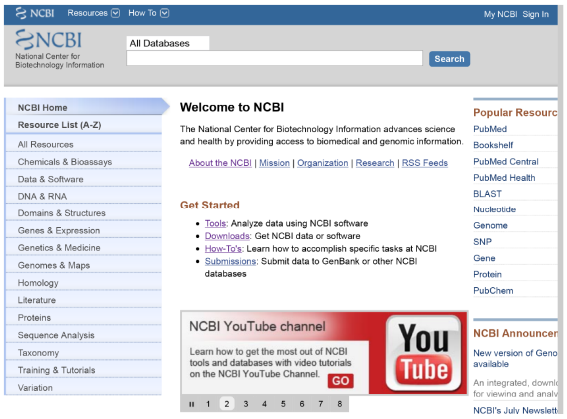
- zahrnují soubory primárních dat – sekvencí DNA a proteinů
  - **Proteinové sekvence**:
    - **PIR,** http://pir.georgetown.edu/
    - **MIPS**, http://www.mips.biochem.mpg.de
    - **SWISS-PROT**, http://www.expasy.org/sprot/

Primární databáze

- Typy sekvencí v primárních databázích
  - Standardní nukleotidové sekvence získané kvalitním sekvencováním
    - **EST**s (**E**xpressed **S**equence **T**ags)
    - **HGTS** (**H**igh **T**hroughput **G**enome **S**equencing)
      - neanotované „surové“ výsledky sekvenačních projektů
    - Referenční sekvence anotovaných genomů
    - **TPA**s (**T**hird **P**arty **A**nnotation)
      - sekvence anotované jinými než původními autory

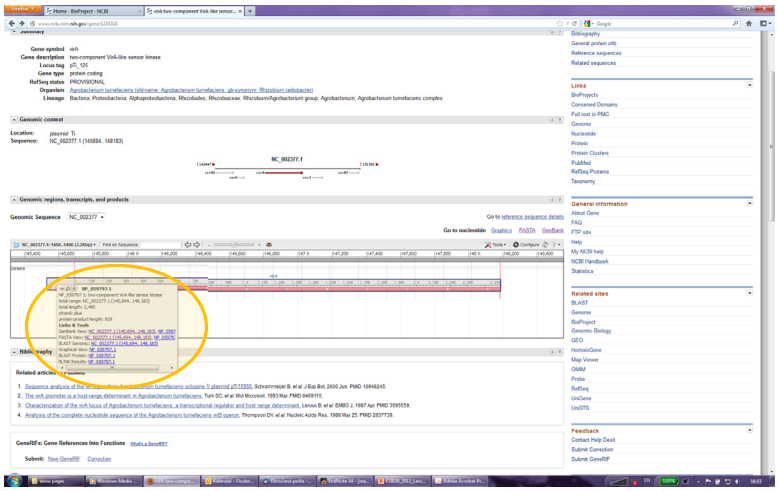
Primární databáze

GenBank (NCBI) https://www.ncbi.nlm.nih.gov/



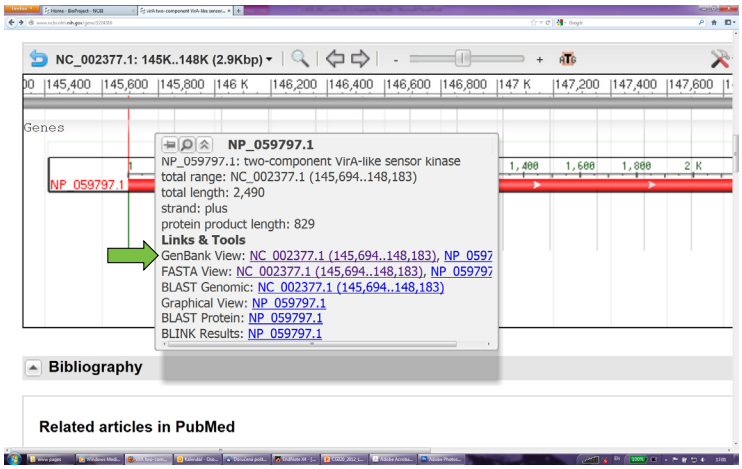
CEITEC

Primární databáze



CEITEC

Primární databáze



Primární databáze

CEITEC

Primární databáze

```
                     /translation="MNGRYSPTRQDFKTGAKPWSILALIVAAMIFAFMAVASWQDNAT
                     TQAILSQLRSINADSASLQRDVLRAHTGTVANYRPIISRLGALRKNLEDLKQLFRQSH
                     IVSESNAAQLLRQLEVSLNSADAAVAAFGAQNVRLQDSLASFTRALSSLPGKASTDQT
                     LEKPTELASMMLQFLRQPSPAISFEISLELERLQKQRGLDEAPVRILAREGPIILSLL
                     PQVKDLVNMIQTSDTAEIAEMLQRECLEVYSLKNVEERSARIFLGSASVGLCLYIITL
                     VYRLRKKTDWLARRLDYEELIKEIGVCFEGEAATTSSAQAALRIIQRFFDADTCALAL
                     VDHDRRWAVETFGAKHPKPVWDDSVLREIVSRTKADERATVFRIISSKKIVHLPLEIP
                     GLSILLAHKSTDKLIAVCSLGYQSYRPRPCQGEIQLLELATACLCHYIDVRRKQTECD
                     VLARRLEHAQRLEAVGTLAGGIAHEFNNILGSILGHAELAQNSVSRTSVTRRYIDYII
                     SSGDRAMLIIDQILTLSRKQERMIKPFSVSELVTEIAPLLRMALPPNIELSFRFDQMQ
                     SVIEGSPLELQQVLINICKNASQAMTANGQIDIIISQAFLPVKKILAHGVMPPGDYVL
                     LSISDNGGGIPEAVLPHIFEPFFTTRARNGGTGLGLASVHGHISAFAGYIDVSSTVGH
                     GTRFDIYLPPSSKEPVNPDSFFGRNKAPRGNGEIVALVEPDDLLREAYEDKIAALGYE
                     PVGFRTFNEIRDWISKGNEADLVMVDQASLPEDQSPNSVDLVLKTASIIIGGNDLKMT
                     LSREDVTRDLYLPKPISSRTMAHAILTKIKT"
ORIGIN
        1 atgaacggaa gatattcacc gacgcggcag gattttaaga caggcgcgaa gccttggtct
       61 atattggccc ttatcgttgc tgcaatgatt ttcgcgttca tggcggttgc gtcctggcag
      121 gacaatgcga ctacccaggc aatcctcagc caactacgat cgattaacgc cgacagcgcc
      181 tcactgcagc gcgatgtact ccgcgctcac acgggcaccg tggcgaacta ccgccccatt
      241 atctccaggc tgggagctct gcggaagaat ctggaagatt tgaagcaatt atttagacaa
      301 tctcatattg taagtgagag caatgctgct caactgctac gccagctaga agtgtctcta
      361 aattcggctg acgcggcggt cgccgccttt ggtgcgcaaa atgtacgcct gcaagattcg
      421 ctggccagtt tcactcgtgc tttgagcagt cttccaggaa aagcctcaac cgatcagact
      481 ttagaaaaac caacagaatt ggctagcatg atgctccaat ttcttcggca accaagcccg
      541 gctatttcat tcgagatcag ccttgaacta gagaggctcc aaaaacaacg cggtcttgat
      601 gaagctcccg tgcgcatact tgcacgtgaa ggtcccatta tcttatcgct tttgccacag
      661 gtgaaagatc tggtgaacat gattcagacg tctgacaccg cagaaattgc ggagatgctg
      721 cagcgcgagt gtttggaggt ctatagcttg aaaaatgtag aggagcggag cgcacgtatc
      781 tttcttgggt ccgcttcagt gggtctttgc ctctacatca tcaccttagt ctataggcta
      841 cgcaaaaaaa ccgattggtt agcgcggcgt ttagattacg aagagctaat caaagagatc
      901 ggagtatgtt ttgaaggtga ggcggccacc acgtcgtccg cgcaagctgc acttcgtatt
      961 attcagcgct tctttgatgc cgatacgtgc gcgttagctc tagtggacca tgaccgtaga
     1021 tgggctgtcg aaacattcgg tgcgaaacac ccaaaacctg tgtgggacga cagcgtgcta
     1081 cgcgaaatag tctctcgtac caaagcggac gaacgggcga cggtattccg catcatatcg
     1141 tcgaaaaaaa tcgtacattt gcctctcgaa attccaggtc tctcgatact actggctcac
     1201 aaatccacag ataaactaat tgcggtttgt tcactgggtt accaaagcta tcgccctcga
     1261 ccttgccaag gcgaaattca gcttcttgaa ctcgccaccg cctgcctctg tcactatatc
     1321 gatgttcggc gtaagcagac cgaatgcgac gttttggcca gacgattgga gcatgcgcaa
     1381 cgccttgagg cagttggtac acttgccggc ggaatagcac atgaatttaa taacatttg
     1441 ggctcaatcc tcgggcacgc agaattagca caaaactcgg tgtctcgaac atctgtcacc
     1501 cgaagatata ttgactatat catttcgtca ggcgacagag ccatgctcat tatcgatcag
     1561 atcttgacgc tgagccgaaa acaggagcgc atgatcaagc catttagtgt ctcagagctt
     1621 gtgaccgaaa tcgctccctt gctacgtatg gctcttccgc caaacatcga gcttagtttc
     1681 agatttgatc aaatgcagag cgtgatcgaa ggaagcccgc ttgaacttca acaggtacta
     1741 attaacatct gcaagaatgc ttcccaagcc atgactgcaa atggtcaaat cgacatcatc
     1801 atcagccaag ctttttacc agttaagaaa attctggcgc atggtgttat gccacctggc
     1861 gactatgttc tcctatctat tagcgacaat ggtggaggca ttcccgaggc tgtgttaccc
     1921 cacattttg aacccttctt tacgacacga gctcgcaacg gtggaacggg tctcggcctt
     1981 gcttctgtgc atggtcatat cagcgcgttt gcgggttaca tcgacgttag ttcaactgtt
     2041 gggcatggga cgcgctttga catttatctc cctccgtctt ctaaggaacc cgtaaatcca
     2101 gacagttttt tcggccgcaa taaggcaccg cgtggaaacg gggagattgt ggcacttgtt
     2161 gagcccgatg acctcctgcg ggaggcgtat gaagacaaga tcgccgctct aggatatgag
     2221 ccggtcggtt ttcgtacctt taatgaaatt cgcgattgga tttcaaaagg caatgaagcc
     2281 gatctggtca tggtcgacca agcgtctctt cctgaagatc aaagtcctaa ttccgtggat
     2341 ttagtgctca agaccgcctc catcatcatt ggcggaaatg atctcaaaat gaccctttca
```


CEITEC

What is an Accession Number?

An **accession number** is **label** that **used to identify a sequence**. It is a **string of letters and/or numbers** that corresponds to a **molecular sequence**.

Examples (all for retinol-binding protein, RBP4):

X02775	GenBank genomic DNA sequence	**DNA**
NT_030059	Genomic contig	
Rs7079946	dbSNP (single nucleotide polymorphism)	
N91759.1	An expressed sequence tag (1 of 170)	**RNA**
NM_006744	RefSeq DNA sequence (from a transcript)	
NP_007635	RefSeq protein	**Protein**
AAC02945	GenBank protein	
Q28369	SwissProt protein	
1KT7	Protein Data Bank structure record	

J. Pevsner, http://www.bioinfbook.org/index.php

CEITEC

NCBI's important RefSeq project:
best representative sequences

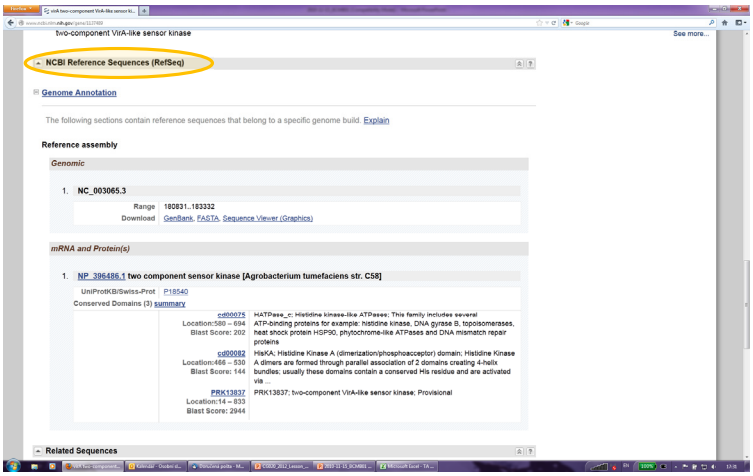
RefSeq (accessible via the main page of NCBI) provides an expertly curated accession number that corresponds to the most stable, agreed-upon "reference" version of a sequence.

RefSeq identifiers include the following formats:

Complete genome	NC_######
Complete chromosome	NC_######
Genomic contig	NT_######
mRNA (DNA format)	NM_###### e.g. NM_006744
Protein	NP_###### e.g. NP_006735

J. Pevsner,
http://www.bioinfbook.org/index.php

RefSeq



CEITEC

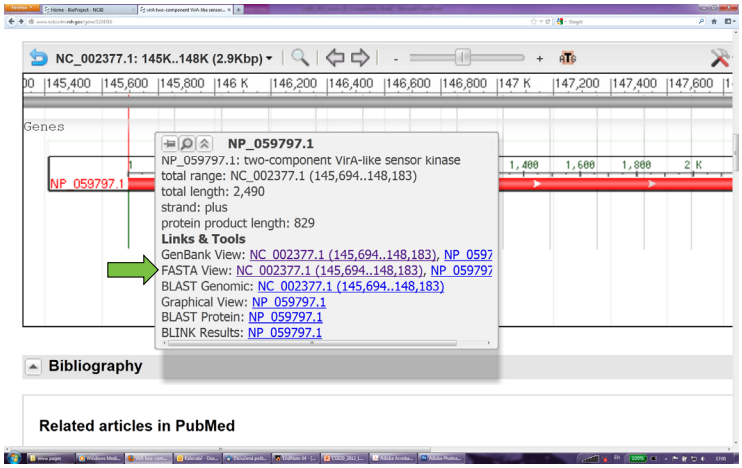
NCBI's RefSeq project: many accession number formats for genomic, mRNA, protein sequences

Accession	Molecule	Method		Note
AC_123456	Genomic	Mixed		Alternate complete genomic
AP_123456	Protein	Mixed		Protein products; alternate
NC_123456	Genomic	Mixed		Complete genomic molecules
NG_123456	Genomic	Mixed		Incomplete genomic regions
NM_123456	mRNA	Mixed		Transcript products; mRNA
NM_123456789	mRNA	Mixed		Transcript products; 9-digit
NP_123456	Protein	Mixed		Protein products;
NP_123456789	Protein	Curation		Protein products; 9-digit
NR_123456	RNA	Mixed		Non-coding transcripts
NT_123456	Genomic	Automated		Genomic assemblies
NW_123456	Genomic	Automated		Genomic assemblies
NZ_ABCD12345678	Genomic	Automated		Whole genome shotgun data
XM_123456	mRNA	Automated		Transcript products
XP_123456	Protein	Automated		Protein products
XR_123456	RNA	Automated		Transcript products
YP_123456	Protein	Auto. & Curated		Protein products
ZP_12345678	Protein	Automated		Protein products

J. Pevsner, http://www.bioinfbook.org/index.php

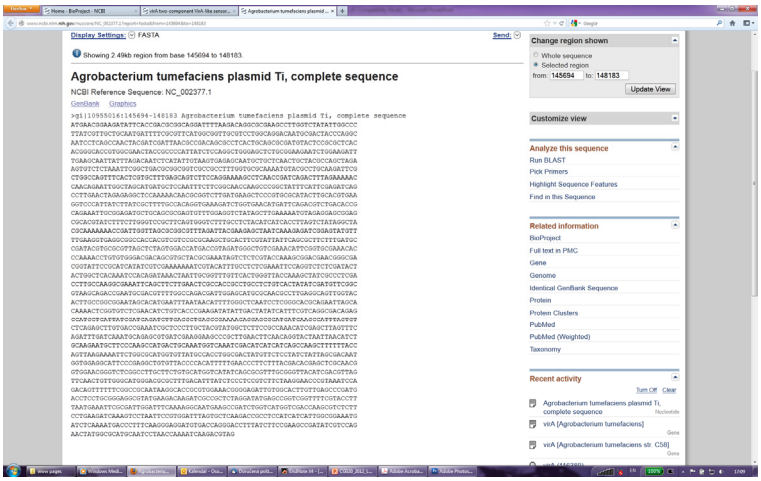
CEITEC

Primární databáze



CEITEC

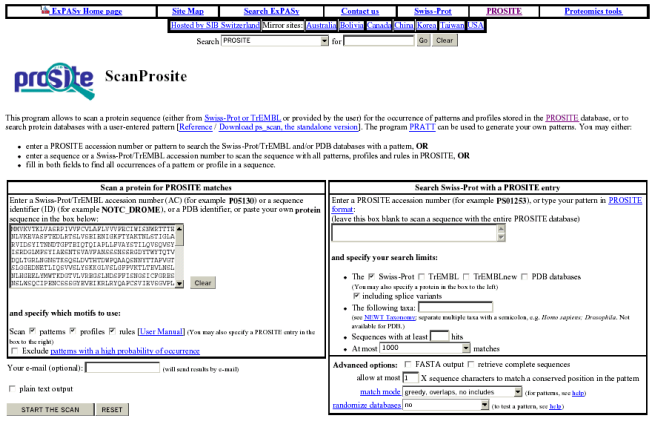
Primární databáze



CEITEC

Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, http://www.expasy.org/prosite/



CEITEC

Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, http://www.expasy.org/prosite/

>PDOC00003 PS00003 **SULFATION** Tyrosine sulfation site [rule] [Warning: rule with a high probability of occurrence].

671 - 685 nkeesntYeteisns

>PDOC00004 PS00004 **CAMP_PHOSPHO_SITE** cAMP- and cGMP-dependent protein kinase phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

744 - 747 SRvT
814 - 817 KRrS

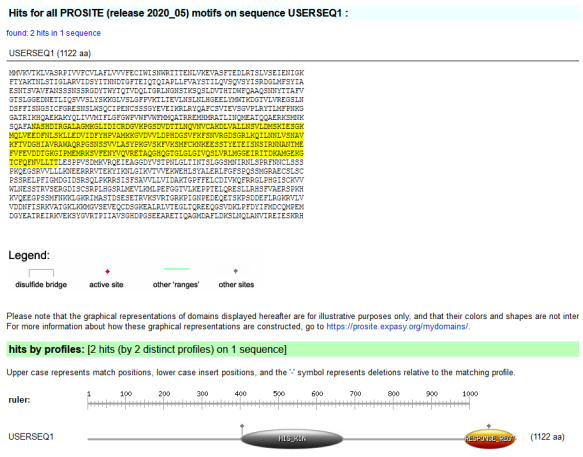
>PDOC00005 PS00005 **PKC_PHOSPHO_SITE** Protein kinase C phosphorylation site [pattern] [Warning: pattern with a high probability of occurrence].

148 - 150 SaR
164 - 166 TgR
171 - 173 SrK
219 - 221 SnK
369 - 371 TvR
460 - 462 SqK
513 - 515 SqR
585 - 587 SiR
602 - 604 TqK
652 - 654 TdK
716 - 718 SpR
726 - 728 SpK
747 - 749 TeK
794 - 796 SaR
854 - 856 ScK
864 - 866 StR
868 - 870 SeR
921 - 923 SpK
967 - 969 SvR
960 - 962 TgR
974 - 976 TsK
997 - 999 SrK
1002 - 1004 TqK
1018 - 1020 SqK
1031 - 1033 TgR
1119 - 1121 SnR

CEITEC

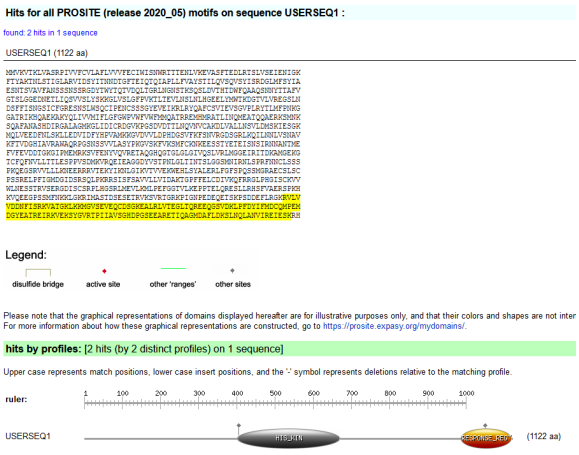
Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, http://www.expasy.org/prosite/



Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
- **PROSITE**, http://www.expasy.org/prosite/



Hits for all PROSITE (release 2020_05) motifs on sequence USERSEQ1 :

found: 2 hits in 1 sequence

USERSEQ1 (1122 aa)

MMVKVIKLVASRPIVVFCVLAFLVVYFECIKSNWRTITENLVKEVASFIEDLRTSLVSEIENIGK
FTVAKTHLSTIGLARVIDSTINNDTGFTEIQTQIAFLFVATSTILQVSQVTSIRDGSLMFSYIA
ESNTSVAVFANSSSNSSRSDVTWYIQTVDQLTGRLNGNSTKSQSLDVTHIDWFQAAQSNNYITAFV
GTSLGGENETLIQGVVSLYSNKKDLVSLGFPVKTLTEVLDSLHLHGEELTWWTKDGTVLVREGSLH
DSFFISNSSICFGRESNSLMSQCIPEHCSSSPYEVEIKRLRVQAFCPVIEVSGVFLRYTLMFPHWG
GATRIKKQAEKAKYQLITVMIFLGFGWPVVWFWMQATRREMNMRATLINQMEATQQAERKSMNK
SQAFAHASDIRGKLAGWGGLIDICRDGVWKSESDVTTTLNQVNVCAKDLVALLNSYLDMSKTESSK
NQLVEEDFNLSKLLEDVIDFYSPVAMGGVDVVLDPRDGSFVKFSNVRGDSGRLKQIINNLVSNAV
KFTVDGSIAVRAWAQRPGSNSSVVLASYPKGVSKFVKSMFCKNWEESSTVETEISNSIRNNAFIME
FYTEVDDTGNGIPWEMKKRVFEKVYQVRETAQHRQGTGLGLGIVQSLVRLANGKEIRIIDKANGEKG
TCFQFNVLLTTLESPPVSDMKVRQEIEAGGDVYSTPKLGLTINTSLGGSMNIRNLSPRFNNCLSSS
PKQEGSKVVLLLKNEERRVTEKYIKNLGIKVTVVEKWEHLSYALERLFGFSPQSNGRAECSLSC
PSSRELPFIGMDGIDSRSQLKWGRSISFSAVYLVYIAACTGFFTLCDIVKQFRRSLFRGISCKVV
WLHESSTRVSERGDISCSRPLHSSRLMEVLKMLPEFGGIVLKEPPTELQRESLLRNSFVAERSPKH
KVQEEGPSSMTNKKLGKRIMASTDSESETRVKSVRIGRKFIGRPEDEQETSKPSDDEFLRGKKVLV
VGDNFISRKVAIGKLKKMGVEVKQCSSGKEALKLVTEGLTGKEQRDVGKLFTYIFNDCQMPEM
DGYEATREIRKVEKSYGVKTPIIAVSGHDPGEEAERETIQAGMDAFLDKSLNQLANVIKEIESKKH

Legend:

disulfide bridge | active site | other 'ranges' | other sites

Please note that the graphical representations of domains displayed hereafter are for illustrative purposes only, and that their colors and shapes are not inter
For more information about how these graphical representations are constructed, go to https://prosite.expasy.org/mydomains/.

hits by profiles: [2 hits (by 2 distinct profiles) on 1 sequence]

Upper case represents match positions, lower case insert positions, and the '-' symbol represents deletions relative to the matching profile.

ruler: 1 100 200 300 400 500 600 700 800 900 1000

USERSEQ1 HISKIN RESPONSE_REGU (1122 aa)

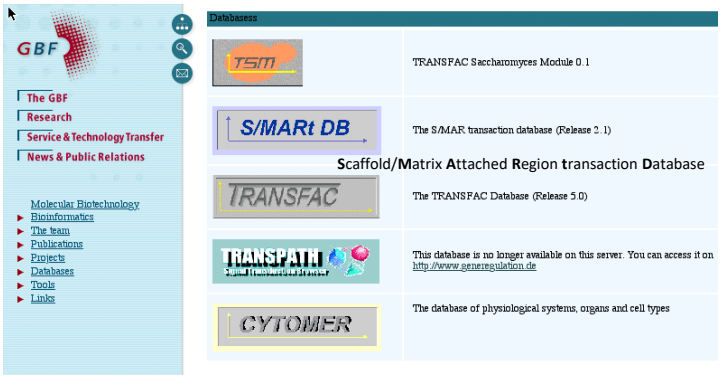
Sekundární databáze

- **Databáze funkčních** nebo **strukturních motivů** získaných srovnáváním primárních dat (sekvencí)
  - **PRINTS**, http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/

CEITEC

Sekundární databáze

- **TRANSFAC** http://www.gene-regulation.com/



CEITEC

S/MARt DB (saffold/matrix attached region transaction database). This database collects information about S/MARs and the nuclear matrix proteins that are supposed be involved in the interaction of these elements with the nuclear matrix. http://transfac.gbf.de/SMARtDB/index.html)

Strukturální databáze

- **PDB** http://www.rcsb.org/pdb/

CEITEC

Strukturální databáze

- **PDB** http://www.rcsb.org/pdb/

Structure Explorer - 1P5Y

PDB
PROTEIN DATA BANK

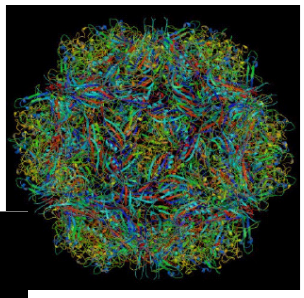
Structure Explorer - 1P5Y

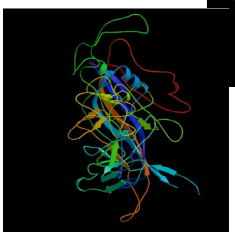
Title **The Structures Of Host Range Controlling Regions Of The Capsids Of Canine and Feline Parvoviruses and Mutants**
Classification **Virus/Viral Protein**
Compound **Mol_Id: 1; Molecule: Coat Protein Vp2; Chain: A; Fragment: Sequence Database Residues 190-737; Engineered: Yes; Mutation: Yes**
Exp. Method **X-ray Diffraction**

View Structure

Summary Information
View Structure
Download/Display File
Structural Neighbors
Geometry
Other Sources
Sequence Details

Explore
SearchLite SearchFields



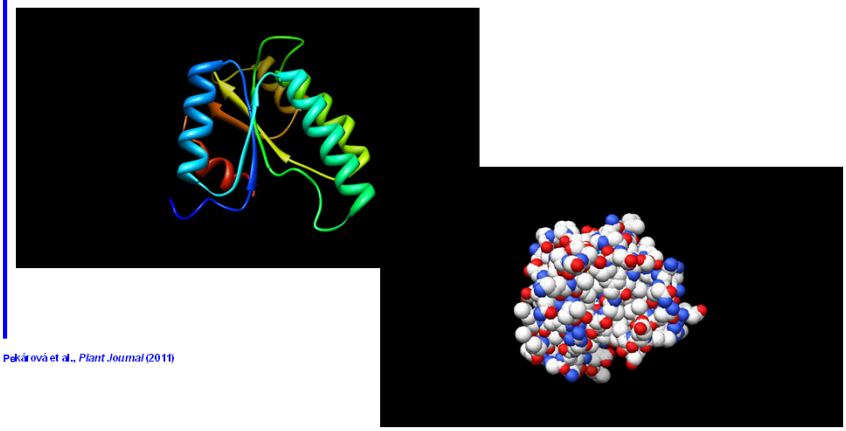


http://www.rcsb.org/pdb/cgi/explore.cgi?job=graphics;pdbId=1P5Y;page=;pid=173561064349344;&bio=1&opt=show&size=500 12/29/2003

CEITEC

Strukturální databáze

- **PDB** http://www.rcsb.org/pdb/



Pekárová et al., *Plant Journal* (2011)

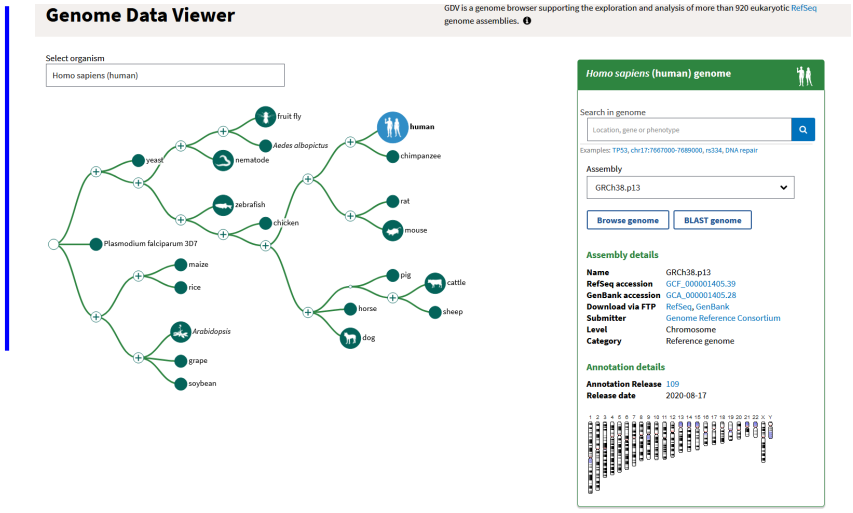
CEITEC

Osnova

- Schéma předmětu
- Definice
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje

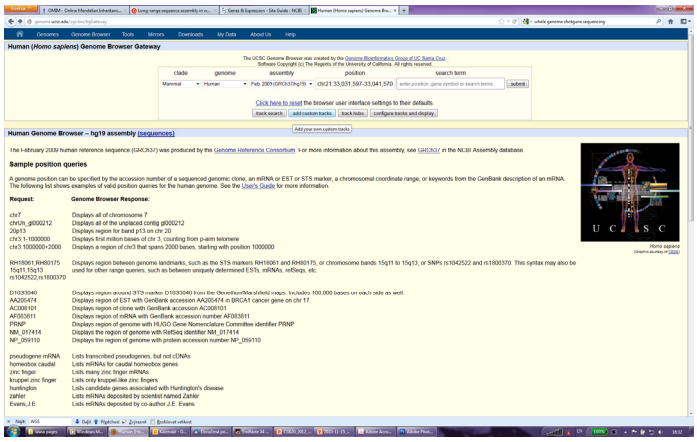
Genomové zdroje

- **NCBI Genome Data Viewer** https://www.ncbi.nlm.nih.gov/genome/gdv /



Genomové zdroje

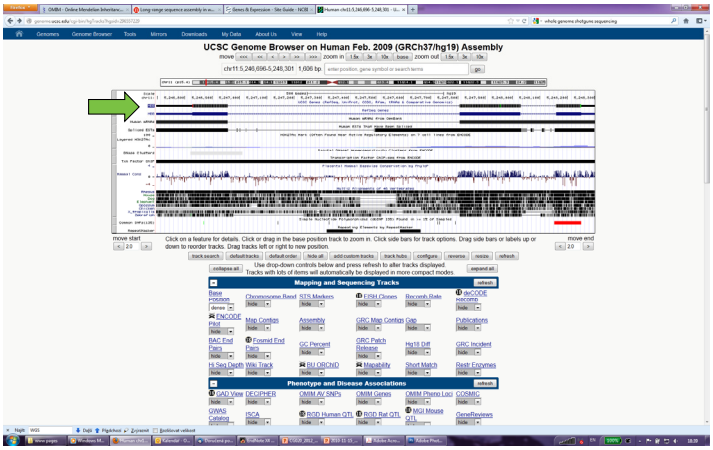
- **Genome Browser Gateway** https://genome.ucsc.edu/



CEITEC

Genomové zdroje

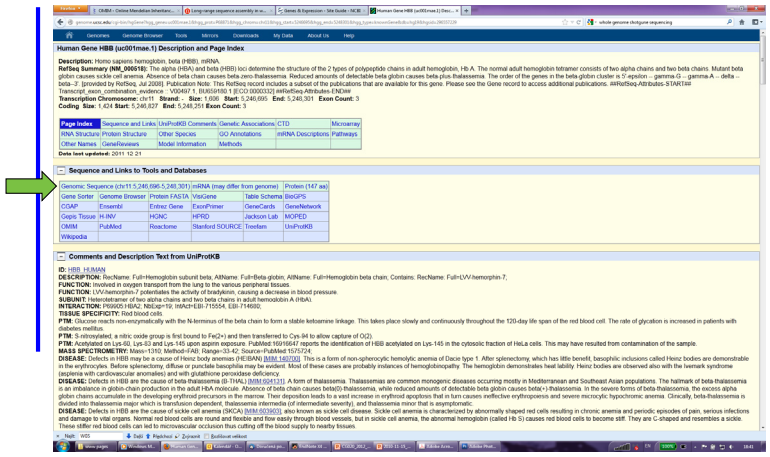
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genomové zdroje

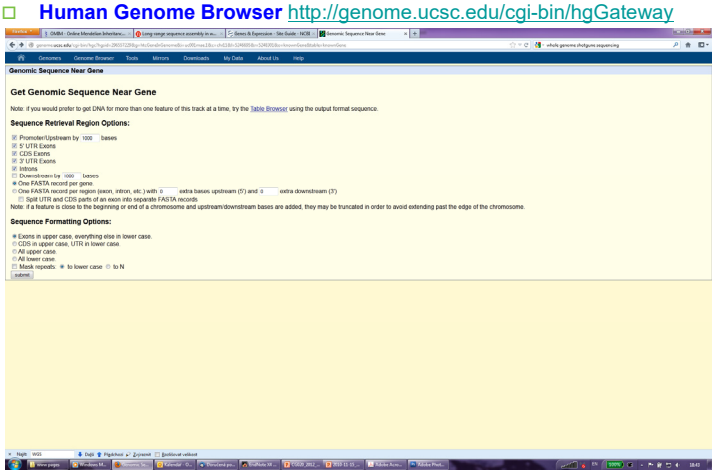
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genomové zdroje

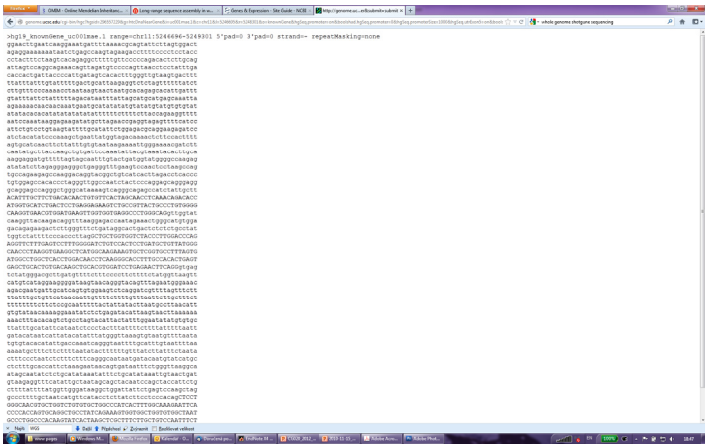
- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genomové zdroje

- **Human Genome Browser** http://genome.ucsc.edu/cgi-bin/hgGateway



CEITEC

Genomové zdroje

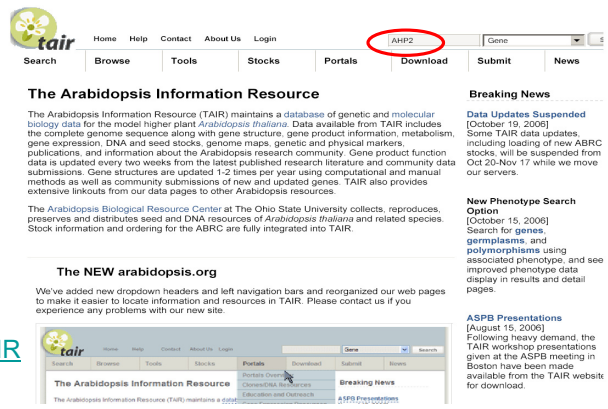
- The Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org

CEITEC

Genomové zdroje

- The Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org

AHP2 @ TAIR



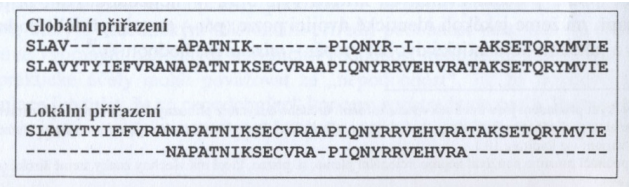
CEITEC

Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií

Analytické nástroje

- Globální vs. lokální přiřazení



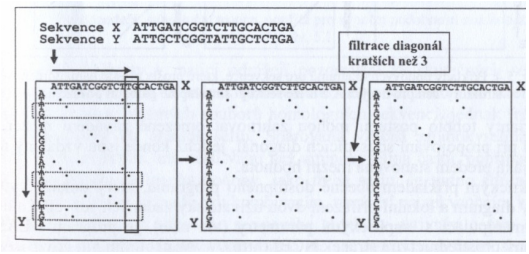
Cvrčková, Úvod do praktické bioinformatiky

- **Globální přiřazení** pouze u sekvencí, které jsou si podobné a podobné délky (za cenu vnášení mezer do jedné nebo obou sekvencí)
- Globální přiřazení se používá především v případě mnohačetného přiřazování (CLUSTALW, viz dále)
- **Lokální přiřazení** umožní identifikaci a srovnání i v případě porovnávání pouze **úseků sekvencí** s významnou mírou podobnosti, např. i při záměně pořadí proteinových domén během evoluce

CEITEC

Analytické nástroje

- Volba správného typu přiřazení pomocí bodového diagramu (dotplot)



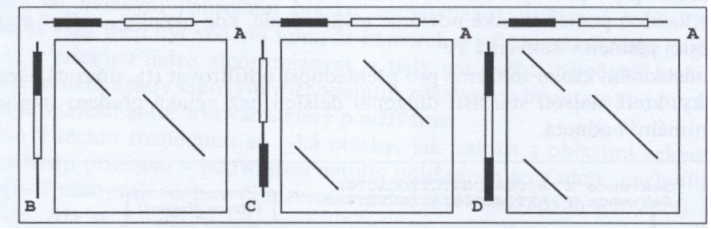
Cvrčková, Úvod do praktické bioinformatiky

  - vynesení sekvencí proti sobě
  - identifikace shody v okně o dané velikosti (např. 2 bp)
  - „odfiltrování" diagonál o délce menší než je mezní hodnota (threshold)

CEITEC

Analytické nástroje

- příklady srovnání sekvencí pomocí bodového diagramu



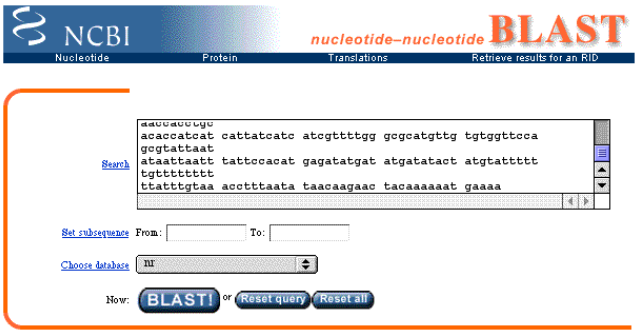
Cvrčková, Úvod do praktické bioinformatiky

- globálně lze srovnávat pouze sekvence A, B
- ostatní sekvence prošly během evoluce záměnou domén a je nutné je porovnávat lokálně
- bodový diagram lze získat pomocí srovnávání programem BLAST2 (viz dále)

CEITEC

Analytické nástroje

- **BLAST** http://ncbi.nlm.nih.gov/BLAST/

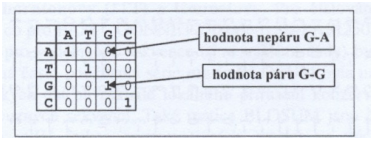


CEITEC

BLAST

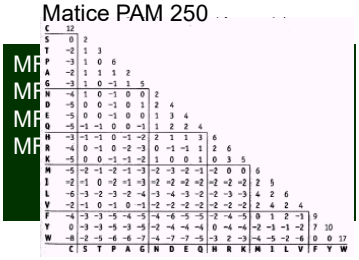
Basic **L**ocal **A**lignment **S**earch **T**ool

- Velikost vyhledávacího slova (word size): 10-11 bp, resp. 2-3 aa
  - Primární podobnosti (seed matches)
  - Rozšiřování oblasti homologie doprava i doleva
- Hodnocení homologie pomocí matice PAM (Point Accepted Mutation) nebo BLOSUM (BLOcks Substitution Matrix)
- Zobrazení výsledků



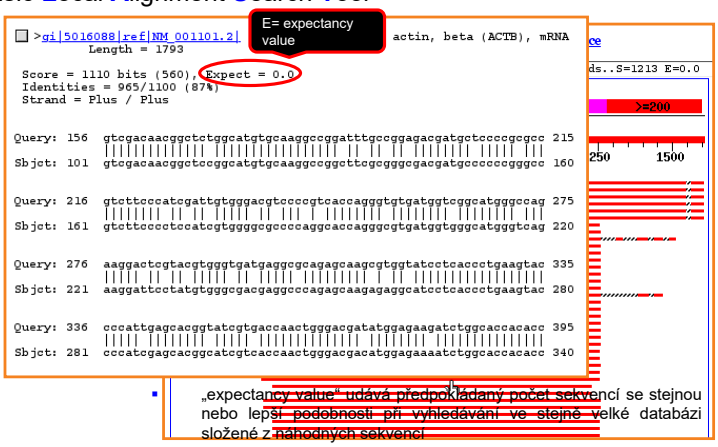
Cvrčková, Úvod do praktické bioinformatiky

Matice PAM 250



BLAST

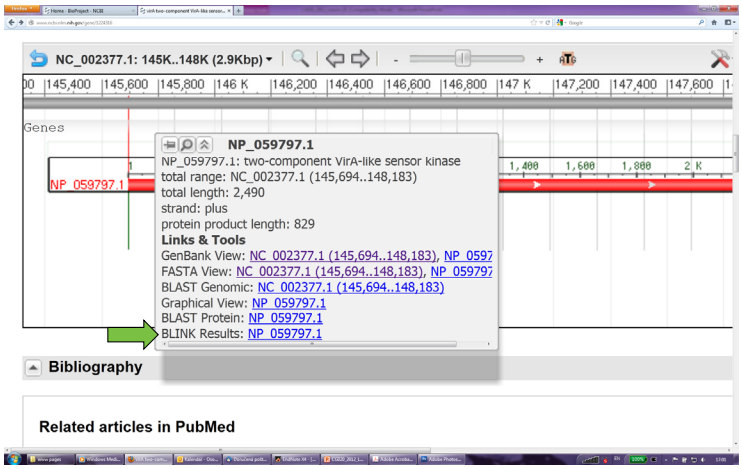
Basic **L**ocal **A**lignment **S**earch **T**ool



- „expectancy value" udává předpokládaný počet sekvencí se stejnou nebo lepší podobností při vyhledávání ve stejné velké databázi složené z náhodných sekvencí
- výsledek udává frakci totožných a u proteinů i podobných pozic, příp. počet vložených mezer

CEITEC

Primární databáze



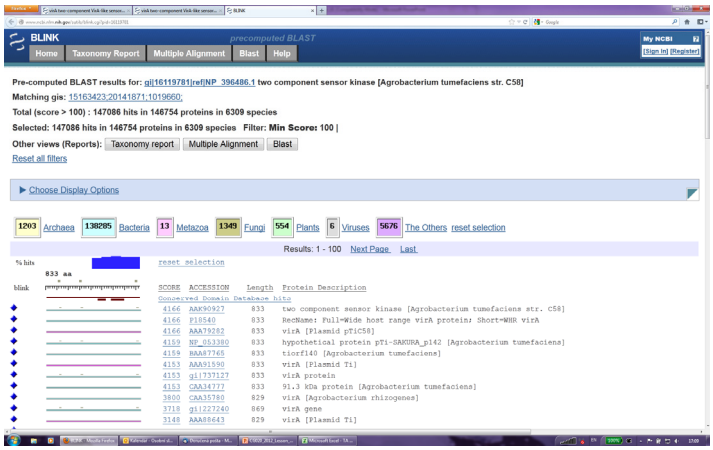
66

CEITEC

BLINK is a link to the pre-computed BLAST search results for the respective sequence (see the next slide).

BLAST

Basic **L**ocal **A**lignment **S**earch **T**ool



CEITEC

BLAST

Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST
  - vyhledávání podle zdroje (organismu) sekvencí, např. známých genomů mikroorganismů
  - **BLASTP**
    - vyhledávání podobnosti k proteinu v databázi proteinových sekvencí
  - **BLASTN**
    - vyhledávání podobnosti k nukleotidové sekvenci v databázi nukleotidových sekvencí
    - další varianty jako např. MEGABLAST pro identifikaci totožných nebo velice podobných sekvencí (vyhledává dlouhé podobné úseky nukl. sekvencí)
  - **BLASTX**
    - vyhledávání podobnosti nukleotidové sekvence přeložené do sekvence aa v proteinové databázi

BLAST

Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST

  - **TBLASTN**
    - vyhledávání podobnosti proteinové sekvence v nukleotidové databázi přeložené do sekvence aa
  - **TBLASTX**
    - vyhledávání k sekvenci nukleotidů přeložené do sekvence aa v databázi nukleotidových sekvencí přeložených do sekvence aa

BLAST

Specializované verze

- V současnosti existuje celá řad specializovaných verzí programu BLAST

  - **PSI-BLAST** (**P**osition-**S**pecific **I**terated **BLAST**)
    - Prvním krokem je standardní BLAST, při kterém PSI-BLAST identifikuje skupinu podobných sekvencí s E hodnotou lepší než minimální hodnota (standardně 0,005)
    - PSI-BLAST vytváří pro každé přiřazení tzv. PSSM (Position Specific Substitution Matrix)
    - PSSM matice zohledňuje výskyt jedné aminokyseliny ve stejné pozici se zvýšenou frekvencí u sekvencí identifikovaných jako podobné v prvním kole pomocí BLAST, což může znamenat funkční konzervovanost

BLAST

Specializované verze

- V současnosti existuje celá řada specializovaných verzí programu BLAST

  - **PHI-BLAST** (**P**attern-**H**it **I**nitiated **BLAST**)
    - Určen k identifikaci specifické sekvence, např. motivu (pattern) v sekvenci podobných proteinových sekvencí
    - Sekvenci motivu je třeba vložit pomocí speciálního syntaxu
      - [LVIMF] znamená buď Leu, Val, Ile, Met nebo Phe
      - - je oddělovník (neznamená nic)
      - x(5) znamená 5 jakýchkoliv aminokyselin
      - x(3, 5) znamená 3 až 5 jakýchkoliv aminokyselin

BLAST

Specializované verze

- Příklad vyhledávání pomocí PHI-BLAST

```
>gi|4758958|ref|NP_004148.1| Human cAMP-dependent protein kinase
MSHIQIPPGLTELLQGYTVEVLRQQPPDLVEFAVEYFTRLREARAPASVLPAATPRQSLGHPPPEPGPDR
VADAKGDSESEEDEDLEVPVPSRFNRRVSVCAETYNPDEEEEDTDPRVIHPKTDEQRCRLQEACKDILLF
KNLDQEQLSQVLDAMFERIVKADEHVIDQGDDGDNFYVIERGTYDILVTKDNQTRSVGQYDNRGSFGELA
LMYNTPRAATIVATSEGSLWGLDRVTFRRIIVKNNAKKRKMFESFIESVPLLKSLEVSERMKIVDVIGEK
IYKDGERIITQGEKADSFYIIESGEVSILIRSRTKSNKDGGNQEVEIARCHKGQYFGELALVTNKPRAAS
AYAVGDVKCLVMDVQAFERLLGPCMDIMKRNISHYEEQLVKMFGSSVDLGNLGQ

[LIVMF]-G-E-x-[GAS]-[LIVM]-x(5,11)-R-[STAQ]-A-x-[LIVMA]-x-[STACV].
```

Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….

Analytické nástroje

https://blog.addgene.org/free-online-molecular-biology-tools

Early Career Researcher Toolbox: Free Online Molecular Biology Tools

By Beth Kenkel

Beth Kenkel
September 12, 2023

Share this article

Primer design. Plasmid mapping. DNA sequence analysis. We all have our favorite tools for tackling these particular tasks, but they tend to be scattered about the internet. To help you keep your virtual molecular biology toolbox organized, today's post features a list of free online molecular biology tools all in one place.

Plasmid mapping

These tools are for viewing, editing or making plasmid maps, but can also analyze and annotate any DNA sequence.

- SnapGene Viewer: The free SnapGene Viewer is great for looking at plasmid maps and viewing sequencing traces, while the paid version provides more tools for plasmid mapping and design (Figure 1).
- Benchling: While you might think of Benchling as an electronic lab notebook, it also has a suite of molecular biology tools and can make plasmid maps. Free for academic users.
- Serial Cloner: Free desktop-based software for plasmid design and mapping.
- ApE (A plasmid Editor): A free, donation-based plasmid analysis tool including editing, annotating, creating maps, and more. This tool is maintained by M. Wayne Davis from the University of Utah.

SnapGene

https://www.snapgene.com/snapgene-viewer/download

CEITEC

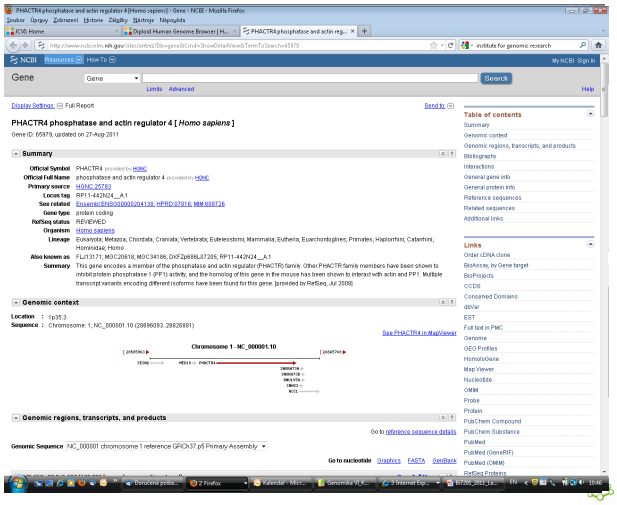
https://www.youtube.com/watch?v=0sQh2s182WQ

Osnova

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

Další WWW zdroje

- TIGR (**T**he **I**nstitute for **G**enomic **R**esearch), http://www.tigr.org/software/
  - Recently part of the J. Craig Venter Institute



CEITEC

Další WWW zdroje

- **O**nline **M**endelian **I**nheritance in **M**an (OMIM) http://www.omim.org/

Shrnutí

- Schéma přednášky
- Role BIOINFORMATIKY v současném pojetí FUNKČNÍ GENOMIKY
- Databáze
  - Spektrum „on-line“ zdrojů
  - PRIMÁRNÍ, SEKUNDÁRNÍ a STRUKTURÁLNÍ databáze
  - GENOMOVÉ zdroje
- Analytické nástroje
  - Vyhledávání homologií
  - Vyhledávání sekvenčních motivů, otevřených čtecích rámců, restrikčních míst….
  - Další www genomové nástroje

Diskuse